Technology Implementation Team Report
2006 April 19

The Technology Implementation Team consists of Ani Au (PIFSC Library), Randy Bossarte (NOAA Miami Library), Donald Collins (NODC), Mary Lou Cumberpatch (NOAA Central Library), Stanley Elswick (NOAA Central Library), Janine Devereaux (Coastal Services Center Library), Brian Voss (NOAA Seattle Library), and Andy Wagliardo (NOAA Central Library). We met in person and via teleconference on a number of occasions since Oct. 2005.

The Technology Implementation Team looked at several technical aspects in its investigation of what system to use for the NOAA Institutional Repository.

Open-source vs. Commercial

We considered the question of whether to purchase a commercial solution or go for an open-source solution or a combination of open-source and home grown customization.

The purchase of a commercial solution has some advantages. It requires less of a investment in staff since the vendor will provide much of the work. The vendor supplies the software and the hardware to run it. Conceivably, NOAA would only have to supply server storage for the documents themselves.

On the other hand, a commercial solution presents several problems. Since most vendors host at least a portion of a repository users would be directed to a site not on a government server. We would need to inform them of that and clear that with NOAA IT security. NOAA may need to hire staff that would have knowledge of the commercial software. This is not as common as that with open-source software which many IT staff are familiar with. At least some vendors also store records in a proprietary format and may or may not provide these records to NOAA in future in a non-proprietary format.

An open-source solution also has distinct advantages and disadvantage. It requires little or no investment of money to implement, it avoids the aforementioned problems with non-governmental sites, and the staffing of administrator and programming positions is easier with open source software. Of the open source solutions we have investigated so far, each uses software and data formats that are non-proprietary. The Library and NODC both have some expertise we can draw on to implement an open-source solution based upon the investigation of systems we have seen so far.

While an open-source solution may cost less money, it will certainly require more staff involvement in the implementation and maintenance.

Also, the pilot project will use the Digital Commons product from proQuest, so we will have some experience with a commercial solution with which to compare the open-source products. So we will focus our attention on open-source software solutions.

We think the advantages outweigh any disadvantages, so we recommend an open-source solution for the institutional repository. We may also recommend some in-house customization. This will depend what we find upon further investigation.

Features of open-source institutional repository software

The basic institutional repository system would allow users to submit documents to the repository via a web-based client interface. It would allow administrators or reviewers to approve the inclusion of these documents in the repository via a client interface. The system would have to allow harvesting of the repository by an OAI-PMH harvester. All of the systems we considered had these basic features.

We needed to investigate more fully the features of a potential IR product to choose one that was appropriate for a NOAA repository. To help us in our quest we consulted the *OSI guide to institutional repository software v. 3.0*. The *Guide* contains a spreadsheet with a number of software packages and the technical details for each. The spreadsheet gave us a great starting point to begin consideration of the features of each system.

The spreadsheet listed the standards each system uses, hardware and supporting software, clients supported, staff skills require to install and configure the system, user registration and authentication details, content submission administration, system-generated statistics and reports, batch importing/exporting features, metadata schemes supported and exported, user interface features, search capabilities, preservation support, version control, and system support and documentation, among others. Each software package had similar capabilities when compared against the list of features in the spreadsheet. In other words, using the spreadsheet did not yield a “clear winner”.

Eventually we cam up with our own list of features that we considered important for our efforts, using some of the ones from the *Guid* and adding our own. We eventually narrowed our list to the following:

- document format types – will the repository software retain the original document format, will it retain only a converted format, or will it retain both (or any number of document formats)?

- deposit structure – where does the software retain the digital objects and can we change that?
- version control – does the software retain all versions of the document? What control do we have over what versions can be made available?
- access control – can the software limit access to a collection within the repository? to a record in the repository? how does it accomplish that? Does the system authenticate users by remote methods like LDAP?
- interface – how easy is the software to use for those who will submit records and those who will search the repository?
- firewalls – how well does the system work within a firewall environment?
- batch loading and exporting – can you batch load? which formats? which formats can you export?
- usage statistics – what statistics can you get from the system? how easily?
- extensibility/interoperability – can you extend the metadata format? how does the system handle admin metadata? preservation metadata? can you harvest from other OAI-PMH repositories? can you allow/disallow harvesting by other OAI harvesters for particular collections or objects?

Narrowing the search

After some initial investigations of the software out there, we eventually decided to take a look at 4 systems. We assigned a person to each system to get more details. We looked at the product web sites and used demonstration systems when available to gain insight into the functionality. We looked at available implementations to see the products “in action”. The four systems we looked at are:

- CDSware. The CERN Document Server Software (CDSware) was developed and is maintained by CERN (European Organization for Nuclear Research) and supports electronic preprint servers, online library catalogs, and other web-based document depository systems. CDSware is unique among the systems looked at in that it maintains its database in the MARC21 format instead of the usual Dublin Core format.
- D-space. D-space was created by MIT as a digital repository to capture the intellectual output of multidisciplinary research organizations. MIT designed the system in collaboration with the Hewlett-Packard Company between March 2000 and November 2002. Version 1.2 of the software was released in April 2004.
- Eprints. The Eprints software has thee largest—and most broadly distributed—installed base of any of the repository software systems described here. Developed at the University of Southampton in England, the first version of the system was publicly released in late 2000.
- Fedora. This system is based on the Flexible Extensible Digital Object and Repository Architecture (Fecora). The system is designed to be a

foundation upon which full-featured institutional repositories and other interoperable web-based digital libraries can be built. Jointly developed by the University of Virginia and Cornell University, the system implements the Fedora architecture, adding utilities that facilitate repository management. The current version of the software provides a repository that can handle one million objects efficiently.

Deeper investigations

In addition to viewing product web sites and online implementation we decided to contact users of each system and make site visits if we could to get actual users' impressions of their systems. We have not yet completed all of our site visits but we did the following so far:

- Fedora. We visited the staff at the National Institute of Standards and Technology (NIST) to discuss Fedora. NIST is also in the initial stages of looking at an institutional repository and has not made a choice, but the person charged with investigating systems favors Fedora since he used it at a previous job at the Library of Congress. He gave us an impressive presentation on Fedora. Fedora is very flexible and full-featured, but requires a lot of expertise to develop since it does not have a presentation layer (you have to build your own). Others have developed open-source and commercial interfaces to Fedora, so perhaps that is not as much of an issue.
- CDSware. CDSware is a very impressive system that comes with the ability to do much more than a repository (CERN uses it for its library catalog in addition to a repository sytem) but vey few institutions use it. We could only locate two in North American that use it (and none within driving distance). We did contact a user in Germany via email who had positive things to say, but that site does not use it as an actual repository, so their experiences there were less than fully useful. We may try to contact one of the North American sites to gain more knowledge, but the lack of a user base and the fact that it uses a lot of different supporting software packages argues against its use, at least until we could load it and see if we can use it.
- Eprints. We have not yet made arrangements for a site visit or teleconference with a user of Eprints. We will look for an opportunity.
- D-space. We visited an office of the Smithsonian Institution which has been using D-space for a small repository since December 2005. The person we spoke with was a librarian with some IT skills who solely maintains the system and showed us a number of the features. The software, by his account, is easy to install and maintain. He is willing to help us out if we decide to try D-space.

Next Steps

We met with NODC Informations Systems Management Division (ISMD) staff to determine whether they objected to any of these products on a security or other basis. They had no problems with installing any of these four software packages.

We also obtained a promise of server space from Parmesh Dwivedi, head of ISMD, to load and investigate different systems.

Given our lack of experience with any of these packages, we think we should load one or more of them and get a real picture of how they would work for us.

Given this same lack of experience, we decided to load D-space in order to:

- gain some knowledge of what is involved in setting up and maintaining a repository
- produce a working product that we can show to those in NOAA who might be able to support a repository either financially or with their participation

Our NOAA Central Library systems maintenance staff (Andy Wagliardo) has loaded the software on an NODC server. The Team plans to meet very soon to discuss what to do next. The agenda will include:

- assignment of any work that we can do, e.g., configuration of collections within the system, designing and creating interfaces like submission forms, user records, etc.
- discussion of testing of functions, e.g., submission of documents, configuring collections, batch loading, changing metadata formats, etc.
- a timeline for implementation
- looking at the software packages inlight of other team reports (metadata, e.g.)
- anything else we haven't thought of

We will retain the option to load one or more of the other software packages later on depending upon our experiences with D-space and our need for features that it may not have.